Sydney Brenner

Comma-less Codes

by

F.H.C. Crick, J.S. Griffith and L.E. Orgel
MRC Unit, Cavendish Laboratory, and Department of Theoretical Chemistry, Cambridge, England.

A Note for the RNA Tie Club May, 1956

It cannot be that axioms established by argumentation can suffice for the discovery of new works, since the subtlety of nature is greater many times than the subtlety of argument.

Francis Bacon

It is assumed in one of the more popular template theories of protein synthesis that amino acids are ordered on a nucleic acid strand. Since there are some twenty naturally occurring amino acids and only four different nucleotides and since the sequences of amino acids so far determined indicate few if any restrictions on the possible neighbours of a given amino acid, it seems probable that more than one nucleotide is required to determine an amino acid. It is not our purpose here to discuss the plausibility of this assumption, but rather to consider, in a formal way, one of the problems which it raises.

Since the number of amino acids exceeds sixteen it seems natural to consider first the case in which an amino acid is determined by three bases. The total number of different sequences of three bases is of course sixty four, so that we need to find some reason why the number of amino acids is so much smaller. Various explanations have already been suggested but here we shall derive the magic number, twenty, by a novel argument.

Suppose we have a sequence of nucleotides forming a nucleotide chain and we associate amino acids with randomly chosen, non-overlapping, consecutive trios of bases as in Fig. 1.

... GACCUGCUAGGACUGCCCAGCU ...

gly glu arg ala

Fig. 1.

It is clear that in general we shall not achieve a satisfactory stacking in the template since there will be places where amino acids are separated by a region of the nucleic acid chain containing a number of nucleotides which is not divisible by three. This type of difficulty can be overcome either by insisting that the amino acids are arranged in the template starting at one end, say the left-hand end, and never filling up a given position until that to the left of it is occupied, or by placing restrictions either on the nucleotide sequences or on the mode of association of amino acids with nucleotides. In the absence of any evidence for or against the former hypothesis we shall develop the latter here.

We suppose that there are certain sequences of three nucleotides with which an amino acid can be associated and certain others for which this is not possible. Using the metaphors of coding we say that certain sequences make sense and the others nonsense. Furthermore we suppose that the sequence of nucleotides present in the nucleic acid makes sense everywhere provided we choose the trios of nucleotides correctly, i.e. occupying positions $3n + 1$, $3n + 2$ and $3n + 3$, but always make nonsense otherwise (see fig. 2).

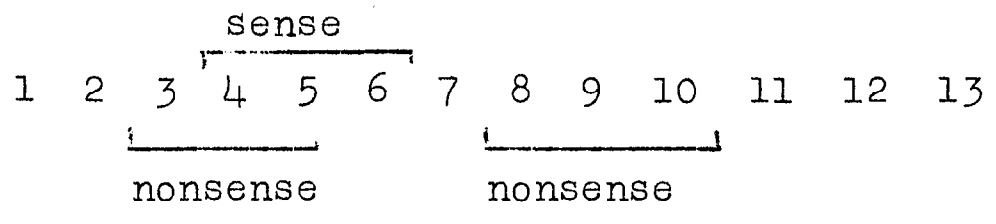

Fig. 2.

Finally we assume that nucleotide sequences corresponding to all possible amino acid sequences are allowed and ask what is the maximum possible number of different amino acids. The nature of this rather abstract problem will be made clear by the following detailed discussion.

(1) Sequence of three identical nucleotides make nonsense.

Proof Suppose AAA were associated with an amino acid α. Then since the amino acid sequence αα is possible so also is the base sequence AAAAAA. However this sequence can be misinterpreted by associating α with the second to fourth, or third to fifth nucleotide.

(2) Not more than twelve sequences of 3 nucleotides, two being identical (for example, AAB) can make sense.

Proof No two such sequences with the same composition (e.g. AAB and ABA) can make sense, for if AAB make sense and is associated with α, then the sequence αα or AABAAB, is possible. This can be misinterpreted unless ABA and BAA make nonse .se.

(3) Not more than eight sequences of three different nucleotides can make sense.

Proof Suppose ABC makes sense and corresponds to α. Then BCA and CAB are forbidden. However one of BAC, CBA, ACB is still allowed In other words not more than two sequences of identical composition are possible, one being an even and the other an odd permutation of the three constituents.

We have thus shown that not more than twenty amino acids are possible in this scheme. That such a number can be achieved we show by construction, one possible selection being

$$A \quad B \quad \begin{matrix} A \\ B \end{matrix} \qquad \begin{matrix} A \\ B \end{matrix} \quad C \quad \begin{matrix} A \\ B \\ C \end{matrix} \qquad \begin{matrix} A \\ B \\ C \end{matrix} \quad D \quad \begin{matrix} A \\ B \\ C \\ D \end{matrix}$$

where $A \quad B \quad \begin{matrix} A \\ B \end{matrix}$ means ABA and ABB. In fact there are ways of choosing the twenty allowed sequences which have a different structure from the above selection.

The problem we have considered is a special case of the more general situation in which one Greek letter is determined by n Roman letters selected from a total of m different Roman letters. One can obtain an upper limit for the number of possible Greek letters in this case by the methods we have used, but it is not in general easy to see whether this upper limit can be achieved. The upper limit of six, corresponding to $n = 2$, $m = 4$, cannot be achieved, only five Greek letters being possible nor can the upper limit be achieved for $n = 2$, $m > 4$. The solution for $n = 3$ and arbitrary m is

$$A \quad B \quad \begin{matrix} A \\ B \end{matrix} \qquad \begin{matrix} A \\ B \end{matrix} \quad C \quad \begin{matrix} A \\ B \\ C \end{matrix} \quad \ldots\ldots\ldots \quad \begin{matrix} A \\ B \\ C \\ J \end{matrix} \quad K \quad \begin{matrix} A \\ B \\ C \\ J \\ K \end{matrix}$$

or more concisely, writing $A_1\ A_2\ \ldots\ A_m$ for the nucleotides, the set $A_i\ A_j\ A_k$ for all $i, j, k = 1,2, \ldots m$, satisfying $k \leqslant j$, $i \leqslant j$. We have not solved the general problem.

Summary

We have shown that while there are sixty four possible sequences of three nucleotides, the requirement that it shall be possible to tell unambiguously which sets of three occupy positions $3n + 1$, $3n + 2$ and $3n + 3$ without reference to the distance from the end of the chain reduces the number of possible sequences which make sense to twenty. It is conceivable that this might have some biological importance, but in any case it makes a nice problem.